Informed and uninformed traders at work: evidence from the French market

Ferriani, Fabrizio

University of Bologna - Department of Economics

18 August 2010

Informed and Uninformed traders at work: Evidence from the French Market

Fabrizio Ferriani*

This draft: August 2010

Abstract

The impact that informed and uninformed agents have on market prices is crucial for informational issues in financial markets. Informed trades are associated with institutional operators while uninformed trades are executed on behalf of retail investors. Using high-frequency data from Euronext Paris, I estimate a model where I take into account traders' identities at transaction level. The results show that when the identities of the traders are different on the two sides of the market, stock prices follow the direction indicated by institutional agents. This means that when the buyer is an informed operator and the seller is a retail one, the former transmits a positive pressure to the market. Conversely, when the seller is an institutional agent and the buyer is an uninformed one market prices depress. There is no significant effect when the agent types are the same on both market sides.

Since traders' identities are concealed in Euronext Paris, the last part of the paper discusses the informational content implicitly provided by observed market variables. Institutional trading is found to increase throughout the day, whereas no evidence of informed trading is found during specific time periods of the continuous auction, except for the first thirty minutes of the day where there are more uninformed trades. Institutional trading is more common during periods of low price changes and high-frequency of transactions. Price variations show that informed agents are usually able to trade at better price conditions. Finally, the tick-test algorithm strongly confirms that informed traders always act as initiators of market transactions.

1 Introduction

The last years have seen an increasing attention on market microstructure as evidenced by the significant amount of contributions to the theoretical and empirical literature. Several factors contributed to it, among of which are the availability of high quality datasets with transaction-level information, the development of new trading systems, the necessity of market regulation and the interest towards market participants' trading strategies. These elements are of primary interest especially for the empirical research devoted to informational concerns in financial markets.

*Dipartimento di Scienze Economiche, Università di Bologna, Piazza Scaravilli, 2, 40126 Bologna, Italy. Email: fabrizio.ferriani@unibo.it.

The empirical features of financial markets have been examined along multiple directions. Hasbrouck (1991) applies VAR models in empirical microstructure to simultaneously model market prices and trade direction. This method proved to be very flexible and has been stretched to take into account the decomposition of informative signals in Hasbrouck (1991), the relationship among integrated markets in Hasbrouck (1995) and the presence of daily time patterns in Dufour and Engle (2000), as well as to test theoretical models in De Jong, Nijman and Roell (1996). Engle and Russell (1998) introduce autoregressive conditional duration models (ACD) to describe the time patterns between consecutive transactions. ACD models received particular interest and have been generalized following the GARCH literature. In Manganelli (2005) it is possible to find a recent contribution that extends ACD models to transaction volumes.

There exists a large portion of empirical microstructure literature that is also devoted to the examination of the presence of informed and uninformed agents in financial markets.Kyle (1985) derives a model where a risk-neutral market maker trades with insiders and liquidity traders. He defines a measure, the Kyle's lambda, that assesses the level of market liquidity as a function of the trading strategies of the two types of market agents. Glostem and Milgrom (1985) propose a sequential model where asymmetric information is directly incorporated into the bid-ask spread: a larger presence of informed agents forces the market operator to widen the interval between the bid and the ask quote. This model is then extended by Easley and O'Hara (1987) to also take into account the order size. Easley et al. (1996) and Easley, Kiefer and O'Hara (1997) introduce in the literature the concept of PIN, a measure of the probability that an informed agent is actually trading in the market. Their model has been widely tested in different markets and it is founded on the idea that a bayesian market maker updates his quotes according to the observed behaviour of market participants. Although I will not take advantage of PIN to evaluate the presence of informed agents in the market, I will exploit the idea that some observed variables can be used to infer the identity of a trader. With reference to the impact that informed traders have on market transactions, Barclay and Warner (1993), Chakravarty (2001) and Alexander and Peterson (2007) investigate the occurrence of stealth trading, i.e. the propensity of informed agents to use medium-sized orders to best exploit their information advantage. The reason of this preference completely stands in volume as information signal. This is because large volumes are easily interpreted by the market as the

attempt of an informed agent to maximize his profits. Meanwhile transactions of small volume are inconvenient because informational advantages expire as long as time passes or because of transaction costs. Foucault, Moinas and Theissen (2007) focus on information asymmetries in a limit order market and analyze the relationship between volatility and bid-ask spread and how the informativeness of the limit order book is affected by the proportion of informed agents in the market. Their study is particularly appealing because they also test the effect of a switch from a fully disclosed market to a regime with hidden identities.

In this paper I extend the ordered probit analysis detailed in Hausman, Lo and Mackinlay (1992), hereafter HLM (1992), to empirically investigate the impact of different market operators on the transactions executed in the Paris Stock Exchange. The aim of this research is twofold: the first is to measure the influence of institutional trading at transaction level by considering the type of agent responsible for eache trade. As it will be discussed, this approach represents a more direct way to measure information asymmetries in the market. The second aim is to study the link between market variables and investors, and to provide a description of the main trading patterns followed by institutional agents.

The paper is organized as follows. Section 2 reviews the model employed by HLM (1992) while Section 3 describes the data used for the empirical analysis. Section 4 details the variables included in the ordered probit specification and how the trader effect is incorporated in the model. Section 5 exhibits the parameter estimates, discusses the marginal effects and shows the robustness tests. Section 6 discusses the information content involved in the observed variables and how this could represent the basis to infer trader identities. Lastly, Section 7 concludes.

2 The Model

This section presents the model used for the empirical analysis, postponing a comprehensive description of regressor specification to Section 4. As anticipated in the Introduction, I consider a generalization of HLM (1992) that takes into account how different kinds of operators affect market transactions. For this reason I will strictly follow the presentation provided in HLM (1992), and I refer for a more exhaustive exposition to the said article. I consider a sequence of transaction prices $P_{t_0}, P_{t_1}, P_{t_2}, ..., P_{t_n}$, observed at time $t_0, t_1, t_2, ..., t_n$, where the generic observation time t_k corresponds to transaction time, i.e. the time between two consecutive

transactions without reference to a fixed sampling frequency. Since the minimum variation in stock prices in Euronext Paris for the selected period is 0.01 euro, the variable tick D_{t_k} is defined as the difference between two consecutive prices multiplied by 100, i.e. $D_{t_k} = (P_{t_k} - P_{t_{k-1}}) * 100$[1]. Hence, D_k represents the multiple of the minimum variation allowed and it provides the price change expressed in euro cents for each couple of transactions. This definition could also be easily extended to other measures of the minimum price variation, such as eighths of dollars for data without decimalization. In the context of the ordered probit model, D_k can be thought as the observed realization of a latent continuous random variable,

$$D_k^* = \mathbf{X}_k'\beta + \epsilon_k \tag{1}$$

where $\mathbf{X}_k$ includes the variables that characterize the mean of D_k^*, and ϵ_k is a Gaussian noise with zero mean and variance equal to $\sigma_k^2 = \mathbf{W}_k'\theta$, where $\mathbf{W}_k$ includes all the variables that affect the variance. In the following I will refer to the variance using the shortcut σ_k^2 postponing the issues relative to variance specification and identification to Section 4. The relationship that links the observed and the latent variable stands in the subsequent interval classification:

$$D_k = \begin{cases} d_1, & \text{if } D_k^* \in A_1 =]-\infty; \alpha_1], \\ d_2, & \text{if } D_k^* \in A_2 =]\alpha_1; \alpha_2], \\ \vdots & \vdots \\ d_m, & \text{if } D_k^* \in A_m =]\alpha_{m-1}; \infty[\end{cases} \tag{2}$$

where $\alpha_1 < \alpha_2 < ... < \alpha_{m-1}$ represent non-overlapping cutpoints that divide the whole data range of D_k^* into m distinct intervals A_j, $j = 1...m$, while d_j defines the outcomes of the observed price change D_k. The issue relative to the choice of the number of intervals m to classify D_k is addressed in Section 4; however, as HLM (1992) emphasized, the choice of m will take into account the trade-off between a decreasing price resolution as long as m increases and the challenges associated with estimating thresholds in extreme intervals collecting only few observations. The assumption of conditional independence and Gaussianity of the error distribution allow to characterize the conditional distribution of D_k in the following way:

[1] In the following, to simplify the notation, I will use only k instead of t_k.

$$P(D_k = d_j|\mathbf{X}_k, \mathbf{W}_k) = \begin{cases} P(\mathbf{X}_k'\beta + \epsilon_k \leq \alpha_1 \mid \mathbf{X}_k) & \text{if } D_k^* \in A_1, \\ P(\alpha_{i-1} < \mathbf{X}_k'\beta + \epsilon_k \leq \alpha_i \mid \mathbf{X}_k) & \text{if } D_k^* \in A_i, 1 < i < m, \\ P(\alpha_{m-1} < \mathbf{X}_k'\beta + \epsilon_k \mid \mathbf{X}_k) & \text{if } D_k^* \in A_m \end{cases} \tag{3}$$

$$P(D_k = d_j|\mathbf{X}_k, \mathbf{W}_k) = \begin{cases} \Phi\left(\frac{\alpha_1 - \mathbf{X}_k'\beta}{\sigma_k}\right) & \text{if } D_k^* \in A_1, \\ \Phi\left(\frac{\alpha_i - \mathbf{X}_k'\beta}{\sigma_k}\right) - \Phi\left(\frac{\alpha_{i-1} - \mathbf{X}_k'\beta}{\sigma_k}\right) & \text{if } D_k^* \in A_i, 1 < i < m, \\ 1 - \Phi\left(\frac{\alpha_{m-1} - \mathbf{X}_k'\beta}{\sigma_k}\right) & \text{if } D_k^* \in A_m \end{cases} \tag{4}$$

where Φ is the standard normal cumulative distribution function. As HLM (1992) stressed, the assumption of normality for the conditional distribution of D_k^* is not mandatory and in principle also other choices are feasible.

I define $\gamma' = [\beta', \theta', \alpha_1, ..., \alpha_{m-1}]$ as the vector of all the parameters included in the model together with D_k^* thresholds; the estimate of γ' is performed with maximum likelihood (ML), given the assumptions on the error distribution. The ML function to be maximized is

$$\sum_{k=1}^{n} \left\{ Y_{1k} \cdot \log \Phi\left(\frac{\alpha_1 - \mathbf{X}_k'\beta}{\sigma_k}\right) + \sum_{i=2}^{m-1} Y_{ik} \cdot \log \left[\Phi\left(\alpha_i - \mathbf{X}_k'\beta\sigma_k\right) - \Phi\left(\frac{\alpha_{i-1} - \mathbf{X}_k'\beta}{\sigma_k}\right) \right] + Y_{mk} \cdot \log \left[1 - \Phi\left(\frac{\alpha_{m-1} - \mathbf{X}_k'\beta}{\sigma_k}\right) \right] \right\} \tag{5}$$

where Y_{ik} is an indicator variable equal to one if D_k belongs to the $i-th$ interval, and equal zero if otherwise.

The theoretical framework detailed above has an analogous specification for interval regression model and can be easily extended to the specified setting with only minor modifications; for a thorough reference, see Cameron and Trivedi (2005) or Wooldridge (2002). However, there are some substantial differences between the two approaches. First, the dependent variable in ordered probit models does not have a quantitative connotation and the interpretation of marginal effects is completely different in the two cases. The estimated coefficients in an interval

regression model directly express the marginal contribution of each variable under the ceteris paribus condition, while the marginal effect in the ordered probit is non-linearly related to the whole set of regressors. This aspect is also reflected in the statistics of interest between the two models, since the interval regression provides a more attractive approach to study the effect that explicative variables have on the conditional mean; conversely, the ordered probit gives a more worthwile emphasis on how regressors influence the conditional probability distribution of D_k^*. A further relevant distinction between these two approaches is the reduced dimension of the γ' vector. In the interval regression, the cutpoints do not enter in the set of parameters to be estimated and are pre-determined by the researcher or by the sampling procedure. While this aspect seems to attribute an advantage for interval regression in terms of immediacy with respect to ordered probit, the last point discussed is far from being a minor issue. Excluding the cutpoints from γ' is a feasible choice only if it is possible to define the range of variation of D_k^* without ambiguity. Moreover, as HLM (1992) emphasized, the estimation of the cutpoints together with the other model parameters allows to fully describe the relationship between the observed realizations and the latent variable. This is also the motivation for the rejection of interval regression as main estimation tool and it is only employed in Section 5 to verify the robustness of the results.

3 The Data

Section 3 describes the data used for the empirical analysis and the variables included in the original dataset, while details relative to variable transformation will be discussed in Section 4. The data have been provided by Eurofidai and they consist of all the transactions registered in the Paris Bourse (Euronext Paris) from 3 February 2008 to 31 March 2008, for stocks belonging to CAC 40, the index that collects the most liquid and traded stocks listed in the Paris Bourse (see Foucault, Moinas and Theissen (2007) for a comprehensive and recent description of the Paris Bourse). All the analyses are focused on the transactions executed during the continuous trading session from 9.00 a.m. to 5.30 p.m.; hence, records relative to the opening and closing auctions have been removed. Even though the opening and closing auctions for some stocks stand for a significative number of observations and for a consistent portion of the daily volume, these two trading sessions are neverthless useless for this research because the said transactions

are executed at the same price. For each stock included in the sample, the variables provided in the dataset refer to actual transactions that have occurred between buyers and sellers during the continuous trading session [2]. The CAC40 stocks have been classified into five groups according to increasing market capitalization on 3 February 2008, the first day of trades included in the dataset. Table 4 displays how stocks have been split into the five categories according to market capitalization quintiles[3]. The first column of table 4 displays the market capitalization in Euro millions for the first day of the sample. The other four columns exhibit some descripitive statistics for the stock sample: the total amount of transactions, the average number of daily transactions, the average transaction volume, and the average price. From the table, it is apparent that there exists quite a variation among the different stocks. This variability is driven by multiple factors, such as the weight of each stock in the CAC40 index, the time period analyzed and company-related events, the interest of operators for a specific stock, and so on.

For each transaction, the dataset displays a code that identifies the operator that takes part in the exchange:

- '1' is the code that refers to transactions executed on behalf of retail investors;
- '2' is the code that refers to transactions executed by operators authorized to trade in the Paris Bourse. They are usually banks or other financial intermediaries, called 'Sociétés de Bourse';
- '6' is the code that refers to transactions executed by the market maker;
- '7' is the code that refers to transactions executed by another kind of financial intermediaries called 'Filiales de la Société de Bourse'. They are financial institutions similar to traders coded with '2'.

It is important to stress that all the transactions are executed by authorized operators, i.e. stock members, but only trades coded with '1' are executed on behalf of retail investors. Trades

[2]The dataset actually registers all the executed transactions. If an order volume on the bid exceeds the available quantity of a corresponding order on the ask, it is recorded only for the executed part. The potential remainder is kept in the order book, without 'walking the book' and can be matched with another order posted by a different agent. In the literature, there is no homogeneity in dealing with order fragmentation and different approaches are proposed (e.g. Chakravarty (2001) or Hasbrouck (1991)). However for the aim of this analysis, orders that are even only partially executed do not represent an obstacle.

[3]Dexia is not included in the sample because of a significant number of transactions coded as executed by the market maker.

classified with '2' or '7' refer to transactions executed by financial intermediaries in their own interest. The sample includes a code for the trader on the two sides of the market, such that the buyer and the seller can represent the same kind of operator or can be different. Since this sample involves highly liquid stocks, there is usually no need for the liquidity provider; hence the number of transactions registered with code '6' is generally absent or extremely limited. All the transactions registered with '1' are attributed to retail or uninformed traders, while all the transactions registered with '2' and '7' are placed in the category of institutional or informed traders. This is a plausible distinction similar to the one proposed by Chakravarty (2001) in his analysis of stealth trading. An important remark is that starting from 23 April 2001, the limit order book has been completely anonymous and traders could no longer view the operator that is actually trading or the category he belongs to. For an empirical analysis of the effects of this anonymous regime, see Foucault, Moinas and Theissen (2007); Section 6 also discusses this issue.

Similar to HLM (1992), I exclude from the final sample all the records where D_k is larger than 35 in absolute value. This procedure, together with the exclusion of transactions classified as executed by the market maker, corresponds to eliminate a very small amount of observations, sometimes even zero. This should not be considered as an arbitrary choice because it does not affect the results but, in turn, reduces the probability of including outliers or mistakes in the reporting data. Stocks with the highest percentage of dropped observations (the maximum is 0.0043% for Alstom) are the ones that also exhibit the highest average price; however this is not surprising because stocks with largest prices are easily subject to a highest average value of D_k with respect to low price stocks. The reason is that if an investor desires to affect the price of a stock by a variation of 0.1%, for instance, the corresponding value of D_k would be larger for high price stocks to produce the same price impact.

Figure 1 shows two frequency plots for one representative stock, Bouygues. There is no particular reason to choose this specific company because in principle all the stocks exhibit quite similar patterns. The choice of focusing on a single stock is only made to save space. The left panel of Figure 1 displays the frequency distribution of the price variation D_k, while the right panel shows the frequency distribution of transaction durations expressed in seconds. With respect to the first graph, there is a noticeable concentration of values for the variable

tick in correspondance of $D_k = 0$. This issue is related to the specificity of the data analyzed. High-frequency financial data report transactions that occur within very short time intervals, which excludes the possibility of large and frequent jumps in prices. Moreover, this aspect is emphasized by the fact that this paper employs highly liquid stocks, where the depth of the limit order book assures execution with limited price skips. As such, the left plot of Figure 1 actually displays a focus of the frequency distribution with a reduced scale; the relative frequency for $D_k = 0$ is explicitly indicated at the top. The data also exhibit features that are quite common in these research topics; similar graphical analyses can be found in HML (1992) or Liesenfeld, Nolte and Pohlmeier (2006). Some stocks exhibit certain peculiarities, wherein the distribution is slightly skewed towards the negative or the positive values. Stocks with higher average prices are characterized by thicker tails and more dispersion, coherent with the previous discussion. Concerning time of transactions, for each trade the data set reports the time at which a transaction occur, expressed as 'hh:mm:ss'. High-frequency data usally display very short durations between consecutive transactions, and this aspect is even more emphasized by the high level of liquidity of CAC40 stocks. Even with synchronous observations at the second level, there is no ambiguity involved as trades are listed in a chronological order. The right plot of Figure 1 shows that most of the trades are executed at the same time or however within very limited time intervals. The graph exhibits a skewness to the left with a noticeable concentration of transactions occurring within zero seconds of time interval. Also in this case the graph provides a focus of the frequency distribution with bold digits for the frequency of zero-second durations. The last interesting aspect to be analyzed involves the examination of the autocorrelogram of D_k, in order to assess the presence of serial correlation in price variation. Figure 2 displays the 30-lag correlogram of D_k , always for Bouygues. From the right graph it is apparent that at least for the first lags the series displays a negative autocorrelation. This is a standard feature that occurs in the whole sample. A possible explanation for this pattern, also observed in other empirical studies, is the one implied by the presence of a bid-ask bounce, with reference to the seminal paper of Roll (1984). However, it is difficult to detect any significant negative autocorrelation for D_k after the fifth lag.

4 The Empirical Specification

This section details the empirical model specification and provides a discussion of the methodology used to include and estimate the trader effect. First, it is necessary to stress that all the stocks exhibit specific peculiarities in terms of concentration of D_k around zero, thickness of tails, number of relevant lags for explicative variables, and so on. However, in the estimation procedure I will try to employ a homogeneous specification for the whole sample as far as possible.

The first empirical issue is related to the number of intervals m, where the simplest solution would be to set the number of intervals equal to the number of all the possible outcomes of D_k. In such a case, for some stocks, e.g. Alcatel, this procedure would generate a limited number of intervals, while for most stocks this choice would produce an excessive number of thresholds to be estimated. Moreover, as it was discussed in Section 2, this brings up some ML estimation concerns about threshold estimates in extreme classes that collect only few observations. Taking this trade-off into account, a feasible strategy to single out the optimal m, consists in being driven by graphical and descriptive analisys. Data aggregation is performed according to the frequency plot discussed in Section 3, and the intervals are set up to replicate data concentration around central values. For the whole stock sample, the number of intervals m is set equal to 5 or 7 or 9, according to the distribution of price variations. Stocks with a higher average price also exhibit a larger frequency of observations in the tails. In the whole sample, a distinct class is reserved to $D_k = 0$ that always includes most of observations. Even if the grouping procedure employed to aggregate data can be thought of as arbitrary, it is worth noticing that, as long as data distribution features are preserved, different interval classifications for D_k involve only slight variations in parameter estimates.

The second issue to be discussed is related to the number of transactions that should be included in the estimation, which depends on the frequency that is used for sampling the data. The possible options are the clock-time convention, where data are selected according to a fixed sampling frequency (such as 5-minute intervals) or the event-time convention where all the transactions are included in the sample after applying the filters discussed above. In their fundamental work relative to the role of time in high-frequency finance, Easley and O'Hara (1992) emphasize how market transactions could be viewed as an optional sampling from the

unobservable continuous process of market price. This aspect complicates the inference as it is impossible to let transaction times be independent and identically distributed. Subsequently, Easley, Kiefer and O'Hara (1997) point out that-clock time stationarity in studies that aim to examine information-based issues could seriously affect the results; a fixed sampling approach implies the loss of the information content included in the time pattern between market transactions. Hasbrouck (2007) emphasizes that the preference for one methodology should be driven by the aim of the research. The main objective of this paper is to highlight the different impacts that informed and uninformed agents have on market trades; hence, the transaction-time approach seem the natural choice as long as transaction time stationarity is assured. This would be consistent with HLM (1992), which use transaction-time events but allow for clock-time effects by including trade durations.

The third and probably more important issue to be examined is the one related to the explanatory variables that will be used in the estimation procedure. This is a crucial point because it implies data manipulation to create regressors that explain the impact of different agents on market transactions. For each observation, I create four new variables that display both the volume of the transaction and the type of operator:

- I generate two dummy variables, 'I' and 'U', that are equal to one if the exchange was executed by an informed or insititutional trader (coded by '2' or '7') or an uniformed or retail trader (coded by '1'), respectively. These variables are defined for both sides of the market in order to specify the type of agent who acts as a buyer and as a seller.

- I multiply the trading volume, and the trader indicators for the two sides of the market for each transaction, in order to obtain four new variables by the matching of the two dummies. In this way it is possible to express together the transaction volume and the types of agents who trade in the market.

As an example, consider the following tables that are obtained from Bnp, where I list the trading volume expressed as the number of stocks exchanged and the two dummy variables that identify the type of trader[4]:

[4]In the empirical analysis, in order to moderate the scale effect implicit in big orders and to avoid excessively small coefficients, the quantity exchanged has been normalized by the average transaction volume computed over the whole sample period. This procedure does not affect the results obtained.

	Volume	Buyer		Seller	
		U	I	U	I
#1	38	1	0	0	1
#2	62	1	0	1	0
#3	113	0	1	1	0
#4	2950	0	1	0	1

Table 1: The table provides four sample consecutive trades for Bnp and the corresponding indicators for the types of agents who executed the transaction.

The first row of Table 1 displays a transaction volume equal to 38, where the buyer is uninformed (U=1) and the seller is informed (I=1). The second one shows an exchanged volume of 62, with the agents uninformed (U=1) on both market sides, and so on. Notice that, as stated before, it is possible to have the same types of operators on both sides of the market. With the matching of all the possible trader combinations according to the procedure outlined above, it is straightforward to obtain the following table:

	Volume	BUSUVol	BUSIVol	BISUVol	BISIVol
#1	38	0.000	38	0.000	0.000
#2	62	62	0.000	0.000	0.000
#3	113	0.000	0.000	113	0.000
#4	2,950	0.000	0.000	0.000	2,950

Table 2: This table indicates the trading volume corresponding to possible trader combinations for each transaction.

where BUSUVol refers to transactions where both the buyer and the seller are uninformed, BUSIVol is for transactions where the buyer is uninformed and the seller is informed, BISUVol is for transactions where the buyer is informed and the seller uninformed, and BISIVol for exchanges where both agents are informed traders. With respect to the first transaction where the buyer is uninformed and the seller is informed, only BUSIVol is different from zero and it is set equal to the exchanged volume registered for that transaction. An analogous conclusion can be easily drawn for the other three scenarios. This procedure is intended to create four new variables that represent four different types of fictitious traders, which explain the trader effect on market transactions. Following this scheme, it is possible to link the information content of the exchanged volume with trader identity for each transaction. It is worth noticing that

for each exchange, there is only one possible combination of buyer and seller that is active, i.e. there is only one possible trader combination that is responsible for the execution of the order, while other types of operators are marked as inactive, with zero volume.

The list of regressors that are included in the mean specification of the ordered probit also involves:

- Time difference between two consecutive transactions (Δt_k) expressed in seconds. All executed transactions are included in the sample, so the aim of this variable is to account for clock-time effects on the conditional mean of D_k^*; relative to this point, a comprehensive explanation could be found in HLM (1992).

- Seasonality. The presence of a seasonal pattern in trades is modelled by a Fourier series with daily periodicity according to this sum:

$$\sum_{i=1}^{p} cos(2\pi i\delta_k) + sin(2\pi i\delta_k).$$

  where δ_k expresses the daily periodicity by the ratio between time elapsed from 9.00 A.M. (i.e the starting point of the continuous auction) and the total duration of each trading day. Daily seasonality is observed especially for volume or volatility patterns, and its presence have been intensely analyzed in the previous research (e.g. Easley and O'Hara (1997)).

- The sign of trade: 'Init'. In the empirical microstructure literature, there are several measures employed to determine the direction of the trade, i.e. to define if a transaction was initiated by a buyer or a seller. For this pourpose the algorithm of reference is the one proposed by Lee and Ready (1991) that classifies transactions according to the so-called 'tick-test'; in this paper I will coherently adopt this procedure as well. [5]

Before exploring ML estimates, there are two more issues that should be examined, i.e. how to exactly specifiy the mean and the variance of the model and the constraints required to achieve a full identification of the vector of parameters γ'. According to the descriptive analysis detailed in

[5] Notice that the data used for the estimation do not include bid-ask quotes and this precludes the possibility of a comparison of the classification obtained using only orders with the one obtained using trade prices. Alternative to tick-test, it is possible to use a simpler classification scheme and to define a variable 'Sign' that is equal +1 if $D_k > 0$, -1 if $D_k < 0$, and 0 if there is no price change. This option does not affect the main outcome of this research.

Section 3, each stock exhibits some peculiarities especially in terms of the distribution of D_k and the magnitude of the exchanged volume. This specificity implies a direct effect on the regressor significance, number of intervals used to classify price variations, optimal number of lagged variables, seasonality, and so on, and is managed according to model parsimony, significance of the parameters, and information criteria. To answer these issues, I have included for each stock four lags of D_k, two lags of 'Init', a $p = 2$ for the seasonal component, and two lags of the four variables created to capture the trader effect on market transactions. Information criteria have been used to choose the best lag specification for the trader effect variables in terms of lags to be included, with possible alternatives between lags 1-2 or lags 2-3[6]. Using a number of lags that is greater than three is difficult to accept even in a high-frequency context; moreover, lags greater than three are often not significant anyway. It is worth stressing that the choice between the two couples of lags is mainly dictated by the better fit of the model indicated by the information criteria. The coefficient estimates are generally similar in terms of direction and significance, even when considering the discarded option.

The last point to be examined in this section is related to identification constraints and is strictly linked to the variance specification adopted. The basic model used as reference employs the mean specification introduced above and a variance normalization. This option corresponds to the absence of an explicit design for the variance σ_k. Without imposing any kind of restriction on the parameters of the model, it is impossible to achieve identification, provided that there exist infinite possible combinations of the parameters β and thresholds $\alpha_1, ...\alpha_{m-1}$ that leave the likelihood unchanged. If an explicit form of heteroschedasticity is not taken into account, identification is achieved by excluding the constant from the list of the regressors, or by fixing a known threshold, α_j. In line with the discussion on interval regression, I have choosen to exclude the constant. Identification constraints slightly complicates whenever one decides to define a set of variables that can affect the variance of the model. The inclusion of a further dimension represented by the scale increases the number of parameter vectors that could generate the same value for the objective function. In this case, identication concerns could be solved by excluding the constant both from the mean and the variance of the error term or by fixing two thresholds. I will return on this topic in Section 5.3 when alternative model specifications are

[6]Contemporaneous values have been excluded because of the endogeneity of the traded volume with the transaction price. A complete discussion on the reasons of this exclusion and a more general treatment of endogenity concerns can be found in HLM (1992).

considered to strenghten the empirical results.

5 The Estimates

5.1 The General Ordered Probit Model

This section examines the ML estimates of the ordered probit model. The ordered probit is employed as the reference model for three main reasons. First it can be easily estimated using any computational or statistical software. Second, it delivers parameter estimates that turn out to be robust with respect to the alternative specifications detailed in Section 5.3. Finally, the maximization procedure and the computation of marginal effects are less time-consuming. Before proceeding to the assessment of the results, it is worth recalling the dimension of the dataset used in this research: Table 4 emphasizes a standard issue in high-frequency datasets, i.e. the number of observations for each stock is extremely large. This point should be carefully considered when analyzing the results, because it becomes easier to obtain significant estimates or, more generally, it is more likely to reject any null hypothesis when the dimension of the dataset increases. This concern has been pointed out by HLM (1992), among others. In line with this, a significance level of 1% (or lower) represents a reasonable choice for hypothesis testing.

The analysis in this subsection is only limited to the model of reference, and Table 5 displays the estimates for two representative stocks, Bouygues and Bnp. There is no particular reason to select these two stocks, except to illustrate the two possible lag alternatives. The complete results for the whole sample are not reported for the sake of brevity, but are available upon request. The findings obtained for Bouygues and Bnp can anyway be generally extended to the full sample. To provide a more complete overview of estimation results, Table 3 summarizes the findings for the variables of interest for all the 39 stocks. Meanwhile, the following outlines the estimation outcomes with respect to each variable:

- At least three of the four lags of D_k included in the mean are negative and statistically significant at 1%. This appears to be a general result that can be extended to the whole sample. These findings are not unexpected; they reflect the pattern displayed by the correlogram of D_k. As it has been highlighted in several studies, this negative pattern is

consistent with the occurrence of reversals in transaction prices.

- The interpretation of the coefficient of Δt_k is not immediate as there is no a homogeneous outcome in the whole sample. This is apparent in Table 5, where Bouygues exhibits a negative coefficient for Δt_k while for Bnp it becomes positive but not statistically significant at 1%. These findings can be extended to the full set of estimates, which also display positive and significant coefficients and seem to perfectly match the ones provided in HLM (1992), where the authors found an ambiguous behaviour for clock-time effects, as well. However, as they pointed out, the absence of a uniform outcome for Δt_k excludes the presence of clock-time effects on the conditional mean of D_k^*, but it does not preclude Δt_k to affect the conditional mean of the observed price changes.

- Table 5 shows that the seasonalilty component does not affect the conditional mean, since the coefficients of the Fourier series are almost never significant at 1%. Using a likelihood ratio test in the whole sample to fully exclude the seasonal component from the conditional mean rejects the null hypothesis only in a limited number of cases (e.g. Bouygues, Vinci, or Crédit Agricole). As anticipated in Section 4, trading seasonality is usually observed for time or volume, since it is largely recognized that these two variables display some clusters during the trading session. In the context of this research, the seasonal component has been included to obtain a more complete model specification.

- The Init variable displays uniformly a negative coefficient for the first lag and either a positive or insignificant coefficient for the second lag. The inclusion of Init in the list of mean regressors should entail the effect of the bid-ask bounce, i.e. it should measure the swinging behaviour of prices between bid and ask quotes. The presence of a second lag for Init that becomes positive or not significant could be due to a decline or a reversal of the autocorrelation pattern already from the second lag. An alternative or perhaps complementary explanation is that the negative pattern entailed in Init is also mainly captured by the lags of D_k. This means that once the effect of past price variations is taken into account, the presence of reversals in buys and sells considerably decreases from the second lag.

- The main focus of the estimation results is on the four variables that describe the trader

impacts on market transactions. From Table 5, a homogeneous outcome for the two stocks is evident, which can be generalized to the whole sample. From the matching of the two different types of traders we obtain the four variables introduced in Section 4. These variables can be categorized as 'cross trading' (BUSIVol and BISUVol), where the operators on the two market sides are different and 'parallel trading' (BUSUVol and BISIVol) where the traders on both market sides are the same. According to results displayed in Table 5, the cross trading is significant at the reference confidence level of 1%, while parallel trading exhibits no significant estimates for both lags [7]. However, the most intriguing findings are in the signs of BUSIVol and BISUVol, which are negative and positive, respectively. The estimates suggest that when an informed agent sells to an uninformed one, he depresses market prices; conversely, an informed agent buying from an uninformed agent transfers a positive pressure to prices. It seems that the market follows the institutional traders' behaviour, but only when the operator type is different on the two market sides. It is worth emphasizing that only cross trading has a significant impact on market transactions, while parallel trading is almost always insignificant, even for BISIVol. If institutional traders are more likely to possess an informational advantage, this asymmetry has an effective gain only when exploited against retail investors. In fact, when there are institutional traders on both market sides, their opposing impacts on market direction are compensated by each other and no significant outcome is produced. Since trader identity is concealed to all market operators, these findings may appear puzzling, at first sight, as they do not provide an immediate way to explain how institutional traders can transfer information to the market. Section 6 discusses how some observed variables could be used to infer traders' identities; any explanation about the transmission of trading pressures to the market is postponed until then. The results relative to cross and parallel trading should also be evaluated in the more general context of information efficiency of financial markets. These findings can be thought of as an ex-post evidence of market efficiency, because the informational advantage of institutional traders is fully incorporated in price changes. Table 3 summarizes how cross and parallel trading effect extend to the whole sample.

[7] It is interesting that cross trading is often still significant even at lower confidence levels.

	0	1	2
BUSUVol	97.44	2.56	0
BUSIVol	2.56	10.26	87.18
BISUVol	7.69	7.69	84.62
BISIVol	89.74	10.26	0

Table 3: This table summarizes the results for the four variables expressing the trader effect. The first column indicates the percentage of stocks where both lags are not significant at 1%, the second column the percentage where at least one lag is significant and finally the third column provides the percentage of stocks where both lags are significant. When significantly different from zero, BUSIVol and BISUVol always display a negative and a positive coefficient, respectively.

Table 3 displays in the first column the percentage of stocks where both lags of the trader effect variable are not significant, the second column provides the percentage where only one lag is significant and the last column indicates the percentage of stocks with both lags that are significant. Obviously, the table is built by considering the case of a negative coefficient for BUSIVol and a positive one for BUSIVol. From the table it is immediate to extend the results discussed in this Section about trader effect to the whole dataset. As for the parallel trading, BUSUVol is never significant at 1% in 97.44% of the sample, while BISIVol is not significant for both lags in 89.74% of the cases; BISIVol exhibits one significant coefficient only for a small fraction of the dataset. Conversely, cross trading displays a strong significance. In fact, more than 80% of the sample exhibits a significant coefficient for BISUVol, and this percentage almost reaches the 90% in the case of BUSIVol[8].

5.2 Diagnostics and Marginal Effects

This section examines some diagnostics about the residuals of the ordered probit, and analyses the marginal effects for the variables that measure the traders' impact on transaction prices: i.e. BUSUVol, BUSIVol, BISUVol and BISIVol. Particular attention is given to the dynamic specification of the model. I will follow the procedure described in HLM (1992), based on the generalized residuals defined in Gourieroux et al. (1985). In a time series context, we expect the model to be correctly specified if residuals do not display serial correlation. Such diagnostics are complicated in the case of latent variable models, because it is impossible to compute residuals,

[8]EdF is the only stock with both lags of BUSIVol not significant (2.5%), Alcatel, Ppr and STM are the three stocks with both lags of BISUVol insignificant (7.69%).

as long as D_k^* is not observed. As such, in the case of ordered probit models, generalized residuals are constructed by exploiting the properties of the normal distribution. More specifically, given that we observe $D_k = d_j$, the generalized residuals $\hat{\epsilon}_k$ can be computed as:

$$\begin{aligned} \hat{\epsilon}_k &= E[\epsilon_k | D_k = d_j, X_k, W_k; \hat{\gamma}] \\ &= \hat{\sigma}_k \frac{\phi(c_1) - \phi(c_2)}{\Phi(c_2) - \Phi(c_1)} \\ c_1 &= \frac{1}{\hat{\sigma}_k} \left(\hat{\alpha}_{j-1} - \mathbf{X}_\mathbf{k}' \hat{\beta} \right) \\ c_2 &= \frac{1}{\hat{\sigma}_k} \left(\hat{\alpha}_j - \mathbf{X}_\mathbf{k}' \hat{\beta} \right) \end{aligned} \tag{6}$$

where $\hat{\gamma}$ is the ML estimation of the parameters, and ϕ and Φ represent the standard normal probability density function and the standard normal cumulative distribution function, respectively. Notice that the previous formula represents a general definition that entails models with an explicit form of heteroschedasticity, although this paper assumes a normalized unit variance for the estimation of the ordered probit model. From this expression, it is straightforward to compute a test that verifies the presence of autocorrelation in the residuals. A full description about how to obtain the score statistics of interest can be found in HLM (1992). The basic idea stands on the fact that under the null hypothesis of no serial correlation, the following score statistics has a χ_1^2 distribution:

$$\hat{\xi}_j = \frac{\left(\sum_{k=j+1}^{n} \hat{D}_{k-j} \hat{\epsilon}_k \right)^2}{\sum_{k=j+1}^{n} \hat{D}_{k-j}^2 \hat{\epsilon}_k^2} \tag{7}$$

$$\hat{D}_k = X_k' \hat{\beta} + \hat{\epsilon}_k. \tag{8}$$

The score statistics can be used to test any order j of serial correlation in the residuals, but it always maintains the same number of degree of freedom, regardless of the value of j. Table 6 displays the values of $\xi_j, j = 1, ..., 8$ for the two stocks Bouygues and Bnp. From the table

it is apparent that all the first four lags of ξ_j are less than 6.6349, the critical value at 1% for a χ^2_1 distribution. After the fourth lag, the behaviour of the score statistics is not as uniform, but the results seem to generally reject the null hypothesis of no serial correlation. This result is not surprising and it completely agrees with the presence of four lags of D_k in the mean specification, as was similarly pointed out in HLM (1992). In considering the whole sample, the absence of autocorrelation cannot be rejected at least for the first four lags for approximately half of the stocks. In some cases it happens to find at least one lag wherein the null hypothesis of no serial correlation is rejected. A similar finding is provided in HLM (1992), even though this research uses a number of observations that is considerably larger. This could be a first explanation for the failure to reject the null hypothesis for some stocks, suggesting that a smaller significance level could be more appropriate. An alternative explanation could be appointed to an excessively limited dynamics for the estimated model. Section 5.3 checks for this possibility: the score statistics is computed in the case of the extended probit model, considering a larger number of lags for all the mean regressors and also non-linear effect for the trader effect variables. Almost half of the sample also continue to fail to reject the null hypothesis for at least one in the first four lags. Moreover, in terms of less autocorrelated residuals, the small benefit that could be observed in some cases, is generally overwhelmed by the loss in terms of model parsimony. Indeed, the extended probit is usually rejected by a LR test with respect to the reference specification. It is worthwhile to mention, for completeness of the exposition, that the rejection of the null hypothesis is particularly evident for Alstom, Lafarge, Unibail and Vallourec. These four stocks have data that were classified into nine intervals, and in this case the probit model delivers generalised residuals that continue to display strong serial correlation.

As far as it concerns marginal effects, for a discrete model like the ordered probit, the quantity of interest for marginal effects is represented by the change in the response probabilities rather than in the expected conditional value (Wooldridge (2002)). This is a direct consequence of using a non-quantitative classification criterion to order the data. The marginal response probabilities can be computed as[9]:

[9]Notice that in the following formula, the regressors do not display the subscript since marginal response probabilities are computed 'at the mean'. The formula refers to the standard probit model, but it can be easily extended to the other specifications.

$$\frac{\partial p(D=d_j|\bar{\mathbf{X}},\hat{\beta})}{\partial X_s}=\begin{cases}-\hat{\beta}_s\phi(\alpha_j-\bar{\mathbf{X}}'\hat{\beta}) & \text{if } j=1,\\ \hat{\beta}_s\phi(\alpha_j-\bar{\mathbf{X}}'\hat{\beta}) & \text{if } j=m,\\ \hat{\beta}_s[\phi(\alpha_{j-1}-\bar{\mathbf{X}}'\hat{\beta})-\phi(\alpha_j-\bar{\mathbf{X}}'\hat{\beta})] & \text{if } 1<j<m\end{cases} \tag{9}$$

where $\bar{\mathbf{X}}$ stands for the vector of means for the explicatives of the model, X_s represents a generical regressor, and $\hat{\beta}_s$ serves as the ML estimate of the corresponding parameter. The marginal response probabilities measure the changes in the probability of observing a specific outcome d_j, for a marginal variation in one of the regressors. In nonlinear models, marginal effects do not correspond to maximum likelihood estimated coefficients, so the interpretation of the impact of each variable is not immediate. This concern is even more consistent in the case of the ordered probit model where the signs of response probabilities for intermediate classes cannot be inferred a priori by the sign of the related coefficient. Only the extreme classes present a direction for the marginal effect that can be directly deduced from ML estimates. Table 8 shows the marginal effects only for Bouygues, but the results can be extended in a similar way to the whole sample and are available upon request. Table 8 consists of seven columns, one for each interval used to classify the values of D_k. The number of marginal response probabilities to be computed clearly depends on the intervals used to partition the distribution frequency of tick. Marginal effects are displayed only with respect to the variables of interest, i.e. the ones that measure the trader effect. The first column shows how the trader effect variables affect the marginal probability for a transaction that belongs to the first interval; analogously, the other columns display the same quantity for the other six intervals. The marginal response probability is not significant for parallel trading across all the seven intervals: this is a direct consequence of the estimation results provided in Table 5, and corresponds to a zero power predictability of BUSUVol and BISIVol on the direction of the trading process. Conversely, the marginal effect for cross trading is significant, but with opposite patterns for BUSIVol and BISUVol. Notice that the outcome relative to the marginal response probability of the central class, including $D_k = 0$, is not unambiguous in the whole sample. In the case of Bouygues, the estimates preserve a significant sign but the magnitude of the effect is considerably reduced. On

the contrary, other stocks exhibit no significant marginal effects for BUSIVol and BISUVol at the central interval. This should not be considered as a shortcoming, but as a decrease of trader capacity to transmit a trend to the market when there is no price change. This is a reasonable finding and it implies an explicit role for price as an informative variable; Section 6 discusses this in more detail. From the inspection of the estimates provided in Table 8, it appears more intriguing to discuss marginal response probabilities of BUSIVol and BISUVol in non-central intervals. To interpret the results, it is essential to remember that the central class always includes zero price variations and it splits the distribution of D_k in negative values on the left and positive values on the right of the central class, respectivley. We see that BUSIVol exhibits a positive sign in the first three classes that collect negative price variations, while the marginal effect becomes positive for intervals that include positive price variations. The opposite pattern is observed in the case of BISUVol, i.e. the marginal response probability is negative when $D_k < 0$, and positive when $D_k > 0$. The appeal of these findings stands in the implication they have with respect to the market direction. When an informed agent has acted as a buyer in the immediate previous transactions, the probability of observing a reduction in current prices is negative, while the probability of observing an increase in prices is positive. Conversely, when the seller is informed, the probability that the current price variation is negative increases, while the probability that $D_k > 0$ reduces. This confirms the analysis discussed in Section 5 and it assigns a positive price pressure to informed traders when they act as buyers, and a negative effect when they trade as sellers. This profile is even more apparent and could be more intuitively realized by looking at Figures 3 and 4. The former plots the marginal response probabilities with respect to BUSUVol and BISIVol for both lags of Bouygues, for all the seven intervals that classify D_k. The blue straight line depicts the marginal response probability, while the two red dashed lines represent confidence intervals at 95%. The latter provides the same graphs for BUSIVol and BISUVol. A graph examination of the four panels in Figure 3 shows how the marginal response probability is never significant, and is fluctuating between quite wide confidence intervals. On the other hand, Figure 4 depicts the behaviour of marginal response probabilities in the case of cross trading, with the opposite swinging path in correspondence of $D_k = 0$ for the two variables. Figure 4 also shows that the marginal effect computed for the class $D_k = 0$ is very close to zero, even if still significant as far as it concerns Bouygues.

With respect to marginal response probabilities it should be recalled that the four variables of interest are expressed with mean-normalized transaction volumes. One could argue that the size of these estimates are noticeably limited. However, this observation should be taken into account along with the fact that these marginal effects could not be extremely large as long as they represent changes in a probability. Moreover, these results are obtained by considering the impact of a marginal variation of only two lags on market prices . Looking at the average number of transaction for each trading day in Table 4, it would be unreasonable to find sizeable effects from the order of some thousands of trades, by contemplating only two lags. Indeed, such a noticeable result should be expected only by time cumulation of a specific trading pattern, like a long occurrence of BUSIVol. However, such would questioneed the constancy of these marginal effects and is out of the scope of this research. Finally, it is worth spending a brief comment on the marginal effect on the conditional mean that are not reported for sake of brevity. In this case, the marginal effect has a unique sign that corresponds to the one of the estimated coefficient. Again, cross-trading has an asymmetric and significant effect on the conditional mean, while parallel trading does not display any significant impact. However, it has to be emphasized that these marginal effects are computed only as a further validation of the results since they do not have a meaningful interpretation in the case of qualitative data.

5.3 Robustness Tests

This section is dedicated to the examination of three alternative model specifications employed to confirm the robustness of the findings described above. The model is re-estimate using an interval regression, an ordered probit with an extended set of regressors and an ordered probit with an explicit form for heteroschedasticity. OLS has been excluded from the set of alternatives used to check the robustness of the results even though it provides an immediate benefit in terms of interpretation of the results and ease of postestimation diagnostics. This choice is driven by two relevant caveats highlighted, among the others, in HLM (1992). First, using a continuous linear model neglects the presence of price discreteness and it forces the dependent variable to assume a continuous attribute. Second, and perhaps more relevant, OLS does not capture nonlinearities implicit in the data, unlike the ordered probit. The interval regression, on the other hand, actually corresponds to an ordered probit model, where the thresholds of data

partition are not estimated (Wooldridge (2002)). With respect to the model of reference, the 'extended' ordered probit includes seven lags for D_k, two lags for Δ_{t_k}, an additional lag for each of the variable that measures the trader effect, and the corresponding squared variables for all the lags considered. This corresponds to the following structure for the mean of a representative stock, where an additional third lag is considered in addition to the standard first two[10]:

$$\begin{aligned} X_k\beta = \sum_{i=1}^{7}\beta_i D_{k-i} + \beta_8\Delta_{t_k} + \beta_9\Delta_{t_{k-1}} + \beta_{10}Init_{k-1} + \beta_{11}Init_{k-2} + \beta_{12}cos(2\pi\delta) + \beta_{13}cos(4\pi\delta) \\ +\beta_{14}sin(2\pi\delta) + \beta_{15}sin(4\pi\delta) + \sum_{i=0}^{2}\beta_{16+i}BUSUVol_{k-i-1} + \sum_{i=0}^{2}\beta_{19+i}BUSIVol_{k-i-1} \\ + \sum_{i=0}^{2}\beta_{22+i}BISUVol_{k-i-1} + \sum_{i=0}^{2}\beta_{25+i}BISIVol_{k-i-1} + \sum_{i=0}^{2}\beta_{28+i}BUSUVol^2_{k-i-1} \\ + \sum_{i=0}^{2}\beta_{31+i}BUSIVol^2_{k-i-1} + \sum_{i=0}^{2}\beta_{34+i}BISUVol^2_{k-i-1} + \sum_{i=0}^{2}\beta_{37+i}BISIVol^2_{k-i-1} \end{aligned} \tag{10}$$

Even if the previous equation implies a considerable number of parameters to be estimated, the crucial point is to validate the results by adding further lags and powers of the variables of interest. If trader effect persists, its impact is expected to still be present even with the inclusion of more regressors. The last alternative employed to assess the empirical findings is the one that entails the following specification for the variance of the error distribution:

$$\begin{aligned} W_k\theta = D_{k-1}\theta_1 + D_{k-2}\theta_2 + \Delta t_k\theta_3 + cos(2\pi\delta_k)\theta_4 + \\ cos(4\pi\delta_k)\theta_5 + sin(2\pi\delta_k)\theta_6 + sin(4\pi\delta_k)\theta_7 \end{aligned} \tag{11}$$

With respect to the model of reference discussed in Section 4, Equation 11 defines an explicit form of heteroschedasticity that introduces a scale factor in the probit estimates[11]. The crucial point relative to this option is to verify that the inclusion of a scale factor should only affect the magnitude of the estimates, keeping the direction of the coefficient constant. To recover empirical features of high-frequency finance, the variance accounts for a seasonal component that evolves through the trading session, which is defined simmetrically to the Fourier series specified

[10] The following formula refers to the case of lags [1,2]. It is immediate to extend it to the case of lags [2,3].

[11] A STATA OGLM routine developed by Richard Williams is employed to achieve maximum likelihood estimates in this last case. The corresponding estimates are labelled as OGLM.

for the mean. The presence of Δt_k should account for clock-time effect in the variance, while lagged price variations control for the magnitude effect related to price changes. The inclusion of the trader effect in the variance has not been considered as preliminary estimates show the absence of clear and unanmbiguous outcome; moreover the estimates are often not significant. Actually, it seems that the asymmetric effect related to trader identity only influences the conditional mean. If the ordered probit model is estimated without imposing a normalization for the variance, identification constraints slightly complicate and require a double restriction on the parameters. As anticipated in Section 4 full model identification is achieved by dropping the constant from both the mean and the variance. Recall that these three alternative specifications are only employed for robustness checks and not as model of reference to assess the presence of a trader effect. This choice is motivated by two main reasons. First, all LR tests and information criteria generally attributes a preference to the ordered probit described in Section 4 beacause of model parsimony. Second, the very high number of observations of sample stocks increases the computation time needed to estimate parameter values and especially the marginal effects, when an explicit form of heteroschedasticity is considered. The findings are however generally left unchanged.

Table 7 shows the estimates for the three specifications depicted in this section, and only for the relevant variables that describe the trader effect. To save space the results refer only to Bouygues and Bnp, but their validity can be extended to the whole sample, and they are available upon request. For both Bouygues and Bnp, the three columns display the parameter estimates for OGLM and for interval regression, and the marginal effect on the conditional mean for the extended probit, respectively. The last point deserves a further explanation because it seems to contradict Section 5.2 about the meaninglessness of marginal effect on the mean. The extended probit model also includes squared variables for trader effect; displaying only the coefficients of first order variables is meaningless, without taking into account the squares. However, the impact of nonlinearities implicit in the squares is considerably smaller, so concavity entails changes in the direction of the marginal effects only for implausible and extremely large values of the exchanged volume. This means that to obtain a minimum level of comparability among the estimates, the use of marginal effect on the conditional mean appears to be a reasonable choice. Even if the estimates cannot be compared in terms of magnitude,

it is apparent from Table 7 that there is uniformity with the results displayed in Table 5 for the standard model. For all the three cases, only cross tarding is significant, with negative sign for BUSIVol and positive sign for BISUVol. Conversely, parallel trading is never significant. Table 7 reinforces the findings discussed in Section 5.1 and it confirms the conclusions drawn from the simpler specification entailed in the standard ordered probit. In the extended probit case, the results about the significance of the parameters generally extends to the further lag included (not displayed in the table). The third or the fourth lag is significant in the case of cross trading, but the rejection of the null hypothesis is not as strong as for the first two lags. This result is in line with the analysis of lag relevance provided in Section 4. Parallel trading appears to be even more insignificant at higher lags, which confirms the hypothesis that there is generally no impact on market prices, when the type of trader is the same on the two market sides. The estimates from OGLM also corroborate the findings obtained in the ordered probit for the mean regressors. As far as it concerns the explicatives included in the variance, the two lags of D_k do not exhibit a homogeneous pattern in the whole sample, and general conclusions about the significance or sign of these variables cannot be drawn. Actually, most of the stocks do not display significant estimates, or have ambiguous direction when they are significant. Conversely, and more interestingly, the time and the seasonality components are generally significant in the whole sample. Time always has a positive sign, similarly to the results discussed in HLM (1992). This implies a positive clock-time effect for the conditional variance of D_k^*, so time elapsing seems to be associated with an increasing variance of the error distribution. The examination of the signs of the coefficients of the terms in the Fourier series is meaningless, but the four terms are almost always significant for the whole sample. This is coherent with the conjecture that daily variance possesses some form of periodicity. LR usually gives the preference for the alternative OGLM model with non-normalized variance, with respect to the original ordered probit. However, the time necessary to estimate the model or the marginal effects using the alternative is considerably longer. Hence, it appears reasonable to prefer the simpler ordered probit model, given that the results are the same, but the estimation time is definitely reduced.

6 Informative Content of Observed Market Variables

Section 5 examined the different impacts of informed and uninformed traders on market transactions, and how this effect only manifests when the type of traders is different on the two market sides. Since traders' identities in Paris Euronext are not available to operators, such that the offers in the limit order book cannot be posted by taking into account which agents are actually filling the book, this section investigates how observed variables can act as signals for the presence of informed operators. Notice that in the followings, I do not claim the detection of the rules behind institutional trading algorithms; the point here is to find some evidence that help to construct a first guess about traders' identities. The list of variables for this analysis is essentially limited, and can mainly be related to exchange durations, intradaily trading pattern, and transaction volumes. Easley and O'Hara (1992), Easley, Kiefer and O'Hara (1997), Foucault, Moinas and Theissen (2007), among others, have highlighted the role of volume, time, price change patterns, bid-ask spread, volatility, and daily periodicity to detect trading by informed investors. All these variables represent public information, as long as they are visible or easily recoverable, by market members from the limit order book. Moreover, the recent diffusion of automated trading algorithms like the volume-weighted average price (VWAP) or the time-weighted average price (TWAP) spreads relevant importance to volume and time as the driving elements for trading[12]. To emphasize how observed variables can convey information on the identity of the traders, I employ the following bivariate probit model:

$$Pr(D_{b_k} = 1|\mathbf{X}_k) = \Phi(\mathbf{X}_k'\beta_b + \epsilon_1) \quad (12)$$

$$Pr(D_{s_k} = 1|\mathbf{X}_k) = \Phi(\mathbf{X}_k'\beta_s + \epsilon_2) \quad (13)$$

$$Cov(\epsilon_1, \epsilon_2) = \rho \quad (14)$$

where D_{b_k} and D_{s_k} are two dummy variables that are equal to one when the trader is informed on the buy side and on the sell side, respectively. With reference to traders'classification proposed in Section 3, $D_{b_k} = 1$ when the buyer is coded with '2' or '7' and $D_{b_k} = 0$ when the buyer is coded with '1'. This sorting immediately extends to the case of sellers, as well.

[12]See Bialkowski, Darolles and Le Föl (2008) or Brownlees, Cipollini and Gallo (2010), for example.

The variables employed as regressors in the bivariate probit are summarized in the following:

- A set of time indicators that identify specific moments of a continuous trading session: D_{open}, D_{lunch}, D_{SP} and D_{clos}. D_{open} identifies a transaction that occurs between 9.00 A.M. and 9.30 A.M., D_{lunch} between 00.30 P.M. and 1.30 P.M., D_{SP} between 3.30 P.M. and 4.00 P.M. and finally D_{clos} between 5.00 P.M. and 5.30 P.M. . These dummy variables identify possible critical periods of a usual trading session. The opening and the closing 30 minutes (D_{open}, D_{clos}) are usually characterized by high volatility, while the lunch time (D_{lunch}) usually features a decrease of trading frequency. Finally, D_{SP} distinguishes the trades occurring within the first 30 minutes from the opening of the NYSE, when trading from institutional investors could be more frequent and sensible to the performance of the U.S. stock market. A variable δ_k that goes from zero to one is also included as a regressor and it is used to model the length of the continuous trading session, as discussed in Section 4.

- The time between consecutive transactions is given by Δt_k. In their seminal work Easley and O'Hara (1992) analyzed the informativeness of market durations and they conclude that a lower trading frequency is usually associated with a lower presence of informed investors in the market. This is due to the lower probability that informed agents trade when an information event has not occurred. This issue has been widely tested, e.g. Dufour and Engle (2000), hence it is reasonable to expect a negative sign for Δt_k. This implies that as long as time elapses the probability of observing $D_{b_k} = 1$ or $D_{s_k} = 1$ decreases.

- The role of volume as information signal is given by: $Volume$ and D_{big}. It is commonly accepted in the literature that institutional or informed traders post larger offers with respect to retail investors. This happens because they want to profit from private information or have constraints to achieve fund performances so they may enter the market with larger orders. On the other hand, the stealth trading literature claims that informed agents employ average-size transactions in order to disguise their presence in the market. I test this interesting hyptohesis in the model by including variables to control for different percentiles of transaction volumes. The results are not displayed because the coefficient of these variables are almost always never significant. This may be due to the absence of

stealth trading in this sample, the presence of some confounding factors that make the identification of stealth trading more difficult, or the absence of a sufficiently large cumulative return required to empirically detect the occurrence of stealth trading. However, the average transaction volumes for all the stocks included in the sample are larger for trades executed by institutional investors; this means that trades with remarkably larger volumes than the average transaction volume could be plausibly interpreted as executed by institutional agents. The variable $Volume$ tries to capture the role of volume by including the number of exchanged stocks for each transaction. During the continuous trading session, volumes usually follow a quite stable pattern even if it is possible to observe an increase in the quantities exchanged at the end of the day. The dummy variable D_{big} is equal to one if a transaction displays a volume that is larger than the average volume of the previous fifteen minutes. The role of D_{big} is to signal if a specific order displays a volume that is larger than the average volume of the immediately previous transactions.

- Squared variation: SV[13]. To construct this variable I have partitioned the trading session into intervals of fifteen minutes each, wherein SV is computed as the squared log difference between transaction prices at the beginning and the end of the 15-minute interval. SV has been included with one lag to avoid simultaneity bias. It is worth noticing that the sum of this squared variation in a day is equal to one of the standard measures of volatility with high-frequency data, i.e. the realized volatility. There exists extensive microstructure literature that deals with high-frequency volatility issues, but a detailed discussion about how to measure volatility overcomes the object of this research. The purpose of SV is far from representing a measure of volatility; it has been included among the regressors to investigate if periods of large price variations have a role in inferring trader identities.

- The price difference at transaction level is provided by D_k, and the indicator that determines which agent initiated the trade is given by $Init$. For each trade, D_k expresses the price variation that has also been employed as the dependent variable in the previous analysis. $Init$ is the variable used to determine the agent that has initiated the transaction, according to the tick-test rule proposed by Lee and Ready (1991).

[13]The variable is multiplied by 100 to obtain coefficients that have, more or less, the same magnitude for all the regressors.

This list only accounts for the observed regressors and can be sorted as time- (D_{open}, D_{lunch}, D_{SP}, D_{clos}, δ_k, and Δt_k), volume- ($Volume$ and D_{big}) or price-related ($Init$, D_k, and SV) variables. To proceed in the examination, each stock is split into two subsamples to produce in-sample and out-of-sample estimates. This is to check if in-sample estimates are helpful in forecasting future trading patterns. More precisely and without loss of generality, the first 80% of the observations are dedicated to achieve in-sample results, which are used to check the performance of the bivariate probit for the other 20% of the observations. The evaluation of forecasts is done according to the quadratic probability score (QPS) defined by Diebold and Rudebusch (1989):

$$QPS = 1/T \sum_{t=1}^{T} 2(P_t - D_t)^2 \tag{15}$$

where P_t represents the bivariate probit probability forecast, and D_t is the corresponding observed realization. The QPS ranges from 0 to 2, with 0 stands as perfect model prediction. This measure has been applied for bivariate probit model by Nyberg (2009). In this specific context, what actually matters to investors is to detect the presence of institutional trading on at least one of the two market sides. This can be done by computing the following two conditional probabilities:

$$\begin{aligned} P_{b_t} &= P_{11_t} + P_{10_t} \\ P_{s_t} &= P_{11_t} + P_{01_t} \end{aligned} \tag{16}$$

P_{b_t} is the conditional probability that informed trading happens for purchases. It is the sum of the two marginal probabilities, P_{11_t}, where the agent is informed on the two market sides, and P_{10_t} where the buyer is informed and the seller is uninformed. It is straightforward to extend this definition to P_{s_t}. The two probability forecasts P_{b_t} and P_{s_t} are employed in Equation 15 to assess the accuracy of the estimates of the bivariate probit.

Table 9 provides parameter estimates of the bivariate probit model for the two representative stocks, Bouygues and Bnp. A discussion of the results limited to these two stocks would not offer an exhaustive analysis, so the following examination is addressed on the basis of Table 10 where the results for the whole sample are summarized. Because of the dimension of the dataset, this case also sees the appropriate significance threshold at 1%. Starting from the four

dummy variables that individuate critical periods during the trading session, Table 9 shows that it is quite difficult to draw a general conclusion on the role of these regressors in signalling the presence of informed agents, except for D_{open}. The variable D_{open} is the only time indicator that shows a quite homogeneous direction, with a negative and statistically significant coefficient for the buy side for almost two-thirds of the sample, and a positive and significant coefficient for 5.13% of the stocks. Similarly, it has a negative sign for the sell side for almost half of the sample, and a positive sign in just more than 20%; in both cases there is almost one-third of the estimates are not significant. These results suggest a strong occurrence of transactions executed on behalf of retail investors during the first 30 minutes of the continuous auctions. This outcome has been found also in Biais, Hillion and Spatt (1995) and Gourieroux, Jasiak and Le Fol (1999). More precisely, Biais, Hillion and Spatt (1995) showed that smaller trades usually occur during the morning, while larger trades are more frequent in the late afternoon. This aspect is attributed to the behaviour of informed agents. The authors suggest as possible explanations the preference of financial intermediaries to trade retail orders at the beginning of the trading day, the tendency to increase trading in the afternoon because of the discovery of the daily fundamental value, or the evaluation of funds with respect to the closing price in the late-afternoon. Conversely, the effect of D_{clos} is split almost equally between informed, uninformed and not significant estimates on both market sides, and the likelihood of observing an informed agent at the end of the day is unclear. If the estimates of D_{lunch} for the buy side are consistent with a large fraction of institutional investors on the market, this effect is less apparent for the sell side, but still reveals a higher presence of informed agents. In any case, the fraction of non-significant estimates is quite high, close or larger than one-third of the sample. The results for D_{lunch} gives the impression that the deduction of Easley and O'Hara (1992) does not hold for transactions executed between 00.30 P.M. and 1.30 P.M. , since the decreasing trading frequency observed at lunch time does not reconcile with a higher fraction of institutional trades. However these findings are not as strong and unambiguous as the ones related to transaction durations that will be examined next. Finally, the outcomes for D_{SP} are different for the two market sides. For D_{s_k}, trades observed immediately after the opening of the US Exchange are coherent with the presence of institutional investors for one-third of the sample, with a large partition of stocks exhibiting insignificant estimates. On the contrary,

in the case of D_{b_k}, transactions executed from 3.30 P.M. and 4.00 P.M. are more likely to be implemented by uninformed investors, even though the intensity is slightly lower.

The four time dummies prove difficult to give inference about trader identities, except for D_{open}. On the other hand, the result relative to δ_k is noticeable. In more than 90% of the sample, the likelihood that a transaction has been executed by an informed investor increases as time elapses during the continuous auction. This result holds for both D_{b_k} and D_{s_k}. The estimate relative to δ_k is completely in line with the work of Biais, Hillion and Spatt (1995) previously discussed, but it is not confirmed for the last 30 minutes of the continuous session, as D_{clos} showed. In any case, this is not necessarily an unappealing outcome, because institutional investors may just be willing to reduce their trading in periods of large price volatility such as the one close to the end of the day.

The estimates concerning trade durations strongly support the thesis of Easley and O'Hara (1992) on time and informed trading. The negative impact of low trading frequency on the probability that a trader is informed is evident from Table 9. This result is extended to the whole sample, where a negative sign is found for almost 75% of the stocks for the buy side and more than the 80% for the sell side. The relationship between time frequency and informational asymmetries has been widely examined starting from Easley and O'Hara (1992), and it is considered one of the more robust ways to detect the presence of informed traders. The estimates of the bivariate probit confirm this theory. These findings hold even with the addition of one lag of this variable, but the strength of this effect becomes less evident.

Volume and D_{big} are the two variables deputed to account for the information content of volume in the bivariate probit. An inspection of Table 10 shows an equivocal result for the role of volume as a signal variable. The estimates for current transaction volume do not exhibit a good degree of predictability about the direction of the effect. Both cases exhibit a high fraction of non-significant coefficients, larger than 30% of the cases. Moreover, it is not possible to clearly state if the current volume has a positive or a negative effect, even if, it would seem the case at least for the sell side. Conversely, the behaviour of D_{big} is quite homogeneous in the whole sample. For the 76.92% of the stocks on the buy side and 84.62% on the sell side, the probability to observe an informed trader increases when the market presents transactions with a larger-than-average volume. These findings may appear misleading at first glance, since they

assign a positive effect to trading volume only when it is related to past volume. However they do not seem so unreasonable, as long as they state that the dimension of the trades per se is not very much informative; it is only when volumes are associated with a time-close reference value, that they are able to unequivocally determine the effect of trader identity.

The two variables $Init$ and D_k should explain any form of price informativeness observed in the transactions. As far as $Init$ is concerned Table 9 displays a positive coefficient for D_{b_k} and a negative coeffcient for D_{s_k}. On the contrary, price variation measured by D_k has a positive effect for purchases and a negative one for sales. These results are very appealing, first of all for their robustness: it is evident from Table 10, that they extend to the whole sample without any exclusion. The interest for this outcome stands, in the ease of its interpretation. $Init$ is the variable that measures which trader initiates the exchange according to the tick-test algorithm. It is equal +1 when the transaction is buyer-initiated, and equal to -1 when it is seller-initiated. The sign of the coefficient of D_{b_k} is positive: this means that when the transaction is buyer initiated the probability that the buyer is an informed agent increases. Conversely, the sign for D_{s_k} is negative: when the transaction is buyer-initiated, it is unlikely that the seller is informed. These findings assign a strong market power to institutional operators and seem to confirm the idea that retail investors act as liquidity traders and the exchanges are caused only by the actions of institutional investors. The conclusion for price variation is specular: D_k has a negative sign on the buy side and a positive sign for the sell side. When the price has augmented with respect to the previous exchange, we expect that the seller is an institutional trader, since it has been able to sell at a larger price. On the contrary, the probability that the buyer is informed increases when the price falls: it corresponds to the case where institutional traders manage to obtain a lower price when they buy. As in the case of lagged durations, these findings are generally confirmed for both market sides also by including a lag of these two variables.

The information content entailed in price movements is also explained by SV, that could be thought of as an indicator for periods of high variability in price movements. It is immediate to recover from Table 9 that this variable has a negative influence on the probability of observing an informed agent which is also validated by Table 10 and holds for more than the 70% of the sample. The negative sign of SV evidences that institutional investors are unlikely to trade

during periods of large price variations, as the previous analysis about D_{open} has indicated. This is not to say that SV could be regarded as a direct measure of volatility, but the strength of this negative effect suggests that informed agents prefer to reduce their trading activities when there are large price movements[14].

Finally, at the bottom of Table 9 it is provided the QPS value for in- and out-of-sample estimates for both market sides. The two values are around 0.50 suggesting a quite appropriate fit, even by using a very simple model like the bivariate probit. These results extend to the whole sample, where I reasonably always find a better prediction for the in-sample than for the out-of-sample case.

7 Conclusions

This paper presents an ex-post analysis of the role of informed and uninformed agents on the market using high-frequency data from the Paris Bourse. Four variables were generated to take into acount the kind of agent responsible for the execution of a trade at transaction level. The results show that when institutional investors are matched with retail investors, they are able to affect market prices. Conversely, there are no significant effects when the agent categories are the same on the two market sides. In particular, an informed buyer is able to transmit a positive pressure on stock prices, while an informed seller succeeds in depressing market prices. These findings are robust to alternative model specifications, and can generally be extended to the stocks that compose the CAC40 index. Because trader identities are concealed in the French Bourse, the outcome may seem quite puzzling, because no one could know who is trading in a specific moment. This also implies that it is not possible to coherently fill the limit order book by inspecting the exact origin of transactions. The last part of the paper examines if observed variables, that are available to market operators, could help to infer traders' identities and to justify the price impacts of insitutional trading. Using a simple bivariate probit, I obtain estimates consistent with the previous literature. The estimates suggest that trading by informed agents are more likely to occur as long as time elapses during the continuous sessions and in periods of high-frequency of transactions. Conversely, they usually avoid the

[14]For sake of completeness it is worth mentioning that the bivariate probit results do not provide a clear and unambiguous explanation of the marginal cases of insignificant estimates for cross trading, as detailed in Section 5.

first half-hour of the continuous auction and more generally periods characterized by large price variations. Informed agents' orders likewise display larger than the average volumes and are executed at better price conditions. Finally, the results emphasize the evident role that informed traders have as the main initiators of market transactions.

References

Admati, A.R., Pfleiderer, P. (1988),"A theory of intraday patterns: volume and price variability", *The Review of Financial Studies*, 1, 3-40.

Alexander, G.J. and Peterson, M.A. (2007), "An analysis of trade-size clustering and its relation to stealth trading", *Journal of Financial Economics*, 84, 435-471.

Barclay, M., Warner, J. (1993),"Stealth trading and volatility: which trades move prices". *Journal of Financial Economics*, 34, 281-305.

Bialkowski, J., Darolles, S. and Le Föl, G. (2008), "Improving VWAP strategies: a dynamic volume approach"*Journal of Banking and Finance*, 32, 1709-1722.

Brownlees, C.T., Cipollini, F. and Gallo, G.M. (2010), "Intra-daily volume modeling and prediction for algorithmic trading", *working paper*.

Biais, B., Glosten, L., Spatt, C. (2005), "Market microstructure: a survey of microfundations, empirical results and policy implications", *Journal of Financial Markets*, 8, 217-264.

Biais, B., Hillion, P., Spatt, C. (1995), "An empirical analysis of the limit order book and the order flow in the Paris Bourse", *Journal of Finance*, 50, 1655-1689.

Cameron, A. C., Trivedi, P. K. (1998), "Regression Analysis of Count Data", *Cambridge University Press*

Cameron, A. C., Trivedi, P. K. (2005), "Microeconometrics methods and applications", *Cambridge University Press*

Chakravarty, S. (2001), "Stealth trading: which traders' trades move stock prices?", *Journal of Financial Economics*, 61, 289-307.

De Jong, F., Nijman, T. and Röell, A. (1996), "Price effects of trading and components of the bid-ask spread on the Paris Bourse", *Journal of Empirical Finance*, 3, 193-213.

De Jong, F., Rindi, B. (2009), "The microstructure of financial markets", *Cambridge University Press*.

Diamond, D.W., Verrecchia, R.E. (1987), "Constraints on short-selling and asset price adjustment to private information",*Journal of Financial Economics* 18, 277-311.

Diebold, F.X., Rudebusch, G.D. (1989), "Scoring the leading indicators", *The Journal of Business*, 62, 369-391.

Dufour, A., Engle, R.F. (2000), "Time and the price impact of a trade", *The Journal of Finance*, 55, 2467-2498.

Easley, D., Kiefer, N.M. and O'Hara, M. (1997), "The information content of the trading process", *Journal of Empirical Finance*, 4, 159-186.

Easley, D., Kiefer, N.M., O'Hara, M. and Paperman, J.B. (1996),"Liquidity, information and infrequently traded stocks", *The Journal of Finance*, 51, 1405-1436.

Easley, D., O'Hara, M. (1987), "Price, trade and information in securities markets", *Journal of Financial Economics*, 19, 69-90.

Easley, D., O'Hara, M. (1992), "Time and the process of security price adjustment", *The Journal of Finance*, 47, 577-605.

Easley, D., O'Hara, M. (1997), "High frequency data in financial markets: issues and applications", *The Journal of Empirical Finance*, 4, 73-114.

Ellis, K., Michael, R., O'Hara, M. (2000), "The accuracy of trade classification rules: evidence from Nasdaq", *The Journal of Financial and Quantitative Analysis*, 35, 529-551.

Engle, R.F., Patton, A.J. (2004), "Impact of trades in an error-correction model of quote prices", *Journal of Financial Markets*, 7, 1-25.

Engle, R.F., Russell, J.R. (1998), " Autoregressive conditional duration: a new model for irregularly spaced transaction data", *Econometrica*, 66, 1127-1162.

Foucault, T., Moinas, S., Theissen, E. (2007), " Does anonymity matter in electronic limit order markets?", *The Review of Financial Studies*, 5, 1707-1747.

Glostem, L., Milgrom, J. (1985), "Bid, ask and transaction prices in a specialist market with heterogeneously informed traders", *Journal of Financial Economics*, 14, 71-100.

Goodhart, C.A.E., O'Hara,M. (1997), "High frequency data in financial markets: issues and applications", *Journal of Empirical Finance*, 4, 73-114.

Gourieroux, Jasiak, Le Fol (1999), "Intra-day market activity", *Journal of Financial Markets*, 2, 193-226.

Gourieroux, C., Monfort, A., Renault, E., Trognon, A. (1987). , "Generalized residuals",*Journal of Econometrics*,34, 1-52.

Hasbrouck, J. (1991), "Measuring the information content of stock trades", *The Journal of Finance*, 46, 179-207.

Hasbrouck, J. (1991), "The summary informativeness of stock trades: an econometric analysis", *The Review of Financial Studies*, 4, 571-595.

Hasbrouck, J. (1995), "One security, many markets: determining the contributions to price discovery", *Journal of Finance*, 50, 1175-99.

Hasbrouck, J. (2007), "Empirical market microstructure", *Oxford University Press*.

Hausman, J., Lo, A. and MacKinlay, A.C. (1992), "An ordered probit analysis of transaction stock prices", *Journal of Financial Economics*, 31, 319-379.

Kyle, A. (1985), "Continuous auctions and insider trading", *Econometrica*, 53, 1315-35.

Lee, C. M. C., Ready, M. J. (1991)"Inferring trade direction from intraday data", *The Journal of Finance*, 46, 733-746.

Liesenfeld, R., Nolte, I. and Pohlmeier, W. (2006), "Modelling financial transaction price movements: a dynamic integer count data model", *Empirical Economics*, 30, 795-825.

Lo, A., MacKinlay, C. and Zhang, J. (2001),"Econometric models of limit-order executions", *Jounrnal of Financial Economics*, 65, 31-71.

Manganelli, S.(2005), "Duration, volume and volatility impact of trades", *Journal of Financial Markets*, 8, 377-399.

Nyberg, H.(2009), "A bivariate autoregressive probit model: predicting U.S. business cycle and growth rate cycle recessions", *working paper*.

Odders-White, E. R. (2000), "On the occurrence and consequences of innacurate trade classification", *Journal of Financial Markets*, 3, 259-286.

Roll, R. (1984),"A simple implicit measure of the bid-ask spread in an efficient market", *Journal of Finance*, 39, 1127-39.

Williams, R. (2009), "Using heterogeneous choice models to compare logit and probit coefficients across group", University of Notre Dame, mimeo.

Wooldridge, J. M. (2002), "Econometric Analysis of Cross Section and Panel Data", *MIT press.*

Table 4: This table classifies the stocks included in the CAC40 index in five groups according to market capitalization quintiles on 3 February 2008. The first column displays market capitalization expressed in Euro millions; the data are obtained from Datastream. The second column exhibits the total number of transactions occured from 3 February 2008 to 31 March 2008. The third column displays the number of average daily transactions. The four and the fifth columns provide average volume and average price, respectively.

	Market Cap	Num. of Trans.	Avg. Trans.	Avg. Vol.	Avg. Pr.
Capgemini	5,245	258,998	6,475	306	35.89
Technip	5,289	181,423	4,536	206	51.02
AF-KLM	5,686	206,466	5,162	469	17.56
STM	6,135	121,228	3,031	2,130	7.66
Lagardere	6,352	122,307	3,058	211	49.53
Vallourec	8,155	330,792	8,270	91	139.81
Alcatel	8,387	205,301	5,133	3,524	3.90
Essilor	8,744	134,672	3,453	232	39.15
Michelin	9,521	296,959	7,424	218	63.38
Accor	10,635	236,324	5,908	265	48.00
Peugeot	11,505	278,311	6,958	308	49.41
Ppr	12,020	175,130	4,378	153	90.85
Eads	12,218	270,256	6,756	704	16.67
Unibail	13,329	189,694	4,742	106	162.10
Bouygues	13,983	274,566	6,864	269	45.72
Pernod	14,301	186,605	4,665	172	69.19
Lafarge	19,009	281,096	7,027	110	151.13
Saint Gobain	19,328	379,089	9,477	274	51.15
Alstom	19,367	324,377	8,109	101	138.57
Renault	19,974	404,396	10,110	241	69.50
Schneider	20,093	346,825	8,671	186	77.65
Veolia	20,786	383,532	9,588	289	51.30
Vinci	22,299	310,050	7,751	276	45.03
Air Liquide	22,795	261,579	6,539	134	93.25
Vivendi	28,827	389,910	9,748	660	25.82
Danone	29,047	355,751	8,894	317	53.67
Crédit Agricole	32,727	399,358	9,978	777	18.45
Carrefour	34,448	308,801	7,720	361	47.13
Lvmh	34,540	310,584	7,764	228	68.57
Société Generale	36,174	803,620	20,090	407	70.80
Gdf	37,623	219,327	5,483	269	37.44
Axa	47,376	537,978	13,449	1,035	21.80
L'Oréal	49,131	279,003	6,975	194	80.39
Suez	54,333	422,244	10,556	435	41.48
France Télécom	55,568	492,229	12,306	1,034	22.47
Bnp	57,864	690,621	17,266	353	60.57
Sanofi-Aventis	64,908	440,472	11,012	439	49.91
Arcelor	75,165	366,871	9,172	568	49.13
Edf	100,419	411,595	10,290	235	63.10
Total	112,685	591,689	14,792	545	48.79

Figure 1: The figure displays the frequency distribution of D_k in the upper plot and the frequency distribution of Δ_{t_k} in the lower plot. The graphs are referred to Bouygues. The two bold numbers indicates the frequency of $D_k = 0$ and $\Delta t_k = 0$, respectively.

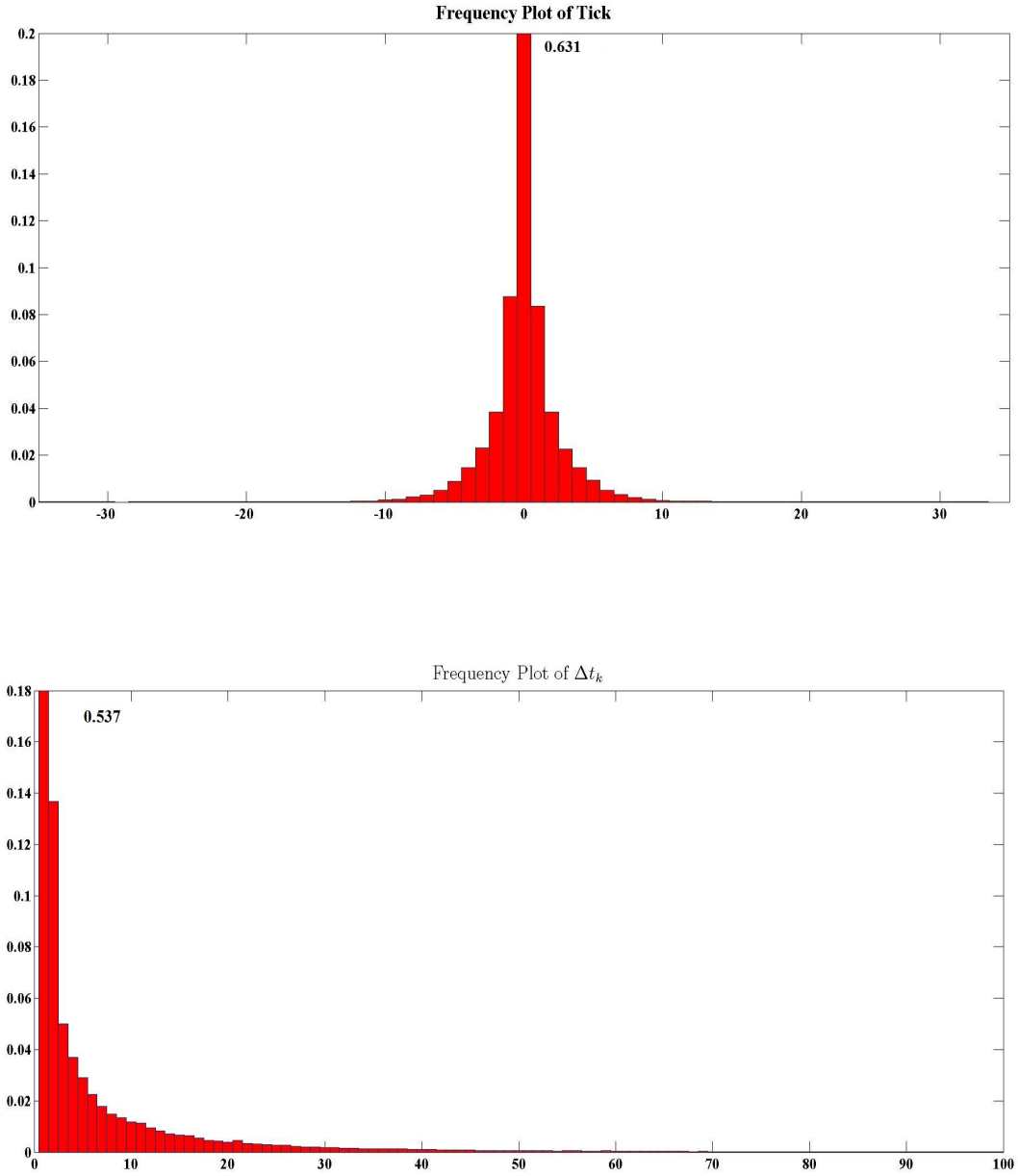

Figure 2: This figure displays the 30-lag autocorrelogram of D_k for Bouygues. The straight lines represent confidence interval at 95%.

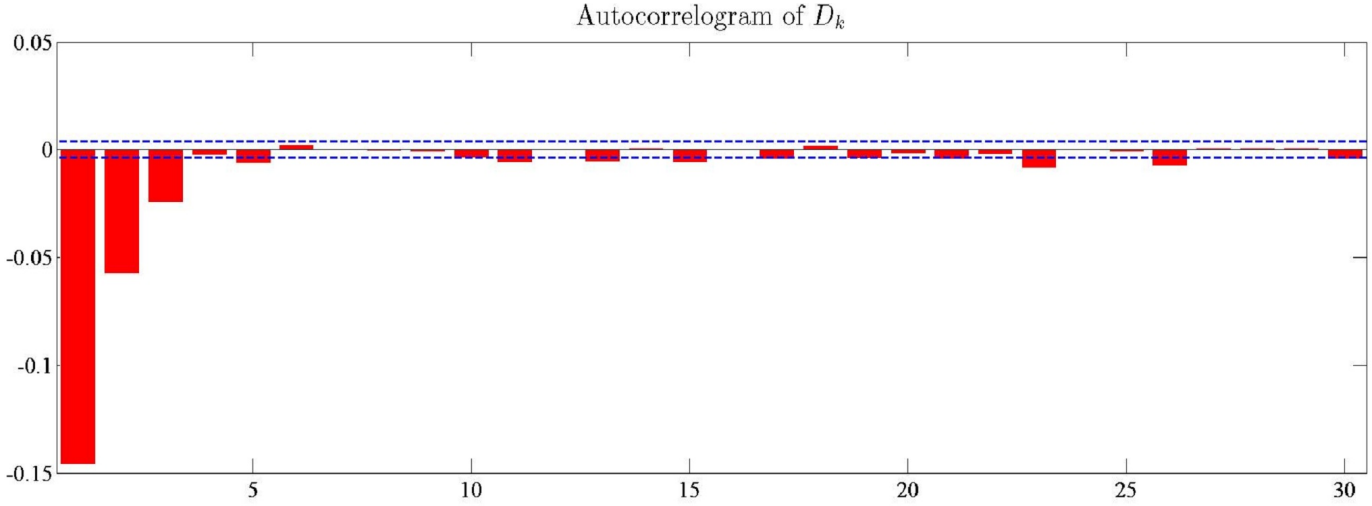

Table 5: This table displays ML estimates for the ordered probit model for two representative stocks, Bouygues and Bnp. The table shows coefficient estimates and P-value in parenthesis. Lags used for the estimation are indicated in brackets.

	Bouygues		Bnp	
Variable	Lags[1,2]	P-values	Lags[2,3]	P-values
D_{t-1}	-4.01e-02	(0.00)	-4.41e-02	(0.00)
D_{t-1}	-3.48e-02	(0.00)	-3.48e-02	(0.00)
D_{t-3}	-1.51e-02	(0.00)	-1.24e-02	(0.00)
D_{t-4}	-4.43e-03	(0.00)	-5.11e-03	(0.00)
Seconds	-7.89e-04	(0.00)	5.78e-04	(0.03)
$Init_{t-1}$	-6.70e-02	(0.00)	-6.90e-02	(0.00)
$Init_{t-1}$	6.79e-02	(0.00)	6.65e-02	(0.00)
$Cos(2\pi\delta)$	-5.44e-03	(0.08)	1.93e-03	(0.31)
$Cos(4\pi\delta)$	-4.23e-03	(0.16)	-1.92e-03	(0.31)
$Sin(2\pi\delta)$	-4.36e-03	(0.15)	6.06e-04	(0.75)
$Sin(4\pi\delta)$	-7.29e-03	(0.00)	-1.91e-03	(0.31)
BUSUVol(t-i)	3.16e-03	(0.11)	4.19e-04	(0.72)
BUSUVol(t-j)	-1.94e-03	(0.35)	-5.25e-04	(0.66)
BUSIVol(t-i)	-1.30e-02	(0.00)	-7.29e-03	(0.00)
BUSIVol(t-j)	-1.64e-02	(0.00)	-8.75e-03	(0.00)
BISUVol(t-i)	2.24e-02	(0.00)	1.20e-02	(0.00)
BISUVol(t-j)	1.72e-02	(0.00)	1.26e-02	(0.00)
BISIVol(t-i)	1.54e-03	(0.44)	1.00e-04	(0.93)
BISIVol(t-j)	-1.71e-03	(0.39)	-6.95e-04	(0.57)

Table 6: This table provides the values of the score statistics, for the null hypothesis of absence of autocorrelation, up to eight lags for Bouygues and Bnp. The critical value for a χ_1^2 at 1% is 6.6349.

Lags	Bouygues	Bnp
ξ_1	0.245	0.925
ξ_2	2.300	5.081
ξ_3	6.105	3.687
ξ_4	4.570	1.456
ξ_5	6.405	19.842
ξ_6	8.388	25.431
ξ_7	17.533	24.996
ξ_8	2.173	44.843

Table 7: This table provides parameter estimates for trader effect variables in the three alternative model specifications. The first column shows the results for an ordered probit model with an explicit form of heteroschedasticity (OGLM). The second column displays the results for interval regression and the third column shows the results for the extended probit model. The table exhibits estimates only for the relevant lags of the trader effect. For the extended probit model marginal effects on the conditional mean are reported. P-values are in parenthesis.

	Bouygues			Bnp		
	OGLM	Interval	Extended	OGLM	Interval	Extended
BUSUVol(t-i)	2.80e-03	3.82e-03	3.50e-03	4.26e-04	9.71e-04	-8.19e-04
	(0.21)	(0.19)	(0.25)	(0.74)	(0.55)	(0.59)
BUSUVol(t-j)	-1.48e-03	-1.93e-03	-3.74e-03	-1.07e-03	-1.39e-03	-6.07e-05
	(0.50)	(0.50)	(0.22)	(0.40)	(0.39)	(0.97)
BUSIVol(t-i)	-1.02e-02	-2.34e-02	-1.57e-02	-5.19e-03	-1.17e-02	-1.03e-02
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BUSIVol(t-j)	-1.60e-02	-2.37e-02	-2.04e-02	-7.81e-03	-1.36e-02	-1.18e-02
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BISUVol(t-i)	1.99e-02	3.66e-02	3.02e-02	9.82e-03	1.90e-02	1.55e-02
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BISUVol(t-j)	1.57e-02	2.50e-02	2.04e-02	1.21e-02	1.97e-02	1.44e-02
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
BISIVol(t-i)	1.49e-03	2.62e-03	4.25e-03	3.33e-04	7.34e-05	2.02e-03
	(0.48)	(0.35)	(0.10)	(0.80)	(0.97)	(0.20)
BISIVol(t-j)	-1.91e-03	-2.61e-03	-2.71e-03	-4.87e-04	-9.64e-04	-1.13e-03
	(0.36)	(0.35)	(0.30)	(0.70)	(0.57)	(0.48)

Table 8: This table displays marginal effects of trader effect variables for Bouygues. For each interval used to classify the data the table provides marginal response probabilities and P-values in parenthesis.

	1		2		3		4	
	Marg.	P-value	Marg.	P-value	Marg.	P-value	Marg.	P-value
BUSUVol(t-1)	-2.51E-04	(0.11)	-2.93E-04	(0.11)	-3.00E-04	(0.11)	1.28E-05	(0.14)
BUSUVol(t-2)	1.54E-04	(0.35)	1.80E-04	(0.35)	1.84E-04	(0.35)	-7.85E-06	(0.36)
BUSIVol(t-1)	1.04E-03	(0.00)	1.21E-03	(0.00)	1.24E-03	(0.00)	-5.29E-05	(0.00)
BUSIVol(t-2)	1.31E-03	(0.00)	1.53E-03	(0.00)	1.56E-03	(0.00)	-6.67E-05	(0.00)
BISUVol(t-1)	-1.78E-03	(0.00)	-2.08E-03	(0.00)	-2.13E-03	(0.00)	9.09E-05	(0.00)
BISUVol(t-2)	-1.36E-03	(0.00)	-1.59E-03	(0.00)	-1.63E-03	(0.00)	6.96E-05	(0.00)
BISIVol(t-1)	-1.22E-04	(0.44)	-1.43E-04	(0.44)	-1.46E-04	(0.44)	6.23E-06	(0.45)
BISIVol(t-2)	1.36E-04	(0.39)	1.59E-04	(0.39)	1.63E-04	(0.39)	-6.93E-06	(0.40)

	5		6		7	
	Marg.	P-value	Marg.	P-value	Marg.	P-value
BUSUVol(t-1)	2.87E-04	(0.11)	2.91E-04	(0.11)	2.53E-04	(0.11)
BUSUVol(t-2)	-1.76E-04	(0.35)	-1.78E-04	(0.35)	-1.55E-04	(0.35)
BUSIVol(t-1)	-1.19E-03	(0.00)	-1.20E-03	(0.00)	-1.05E-03	(0.00)
BUSIVol(t-2)	-1.50E-03	(0.00)	-1.51E-03	(0.00)	-1.32E-03	(0.00)
BISUVol(t-1)	2.04E-03	(0.00)	2.06E-03	(0.00)	1.80E-03	(0.00)
BISUVol(t-2)	1.56E-03	(0.00)	1.58E-03	(0.00)	1.38E-03	(0.00)
BISIVol(t-1)	1.40E-04	(0.44)	1.42E-04	(0.44)	1.23E-04	(0.44)
BISIVol(t-2)	-1.56E-04	(0.39)	-1.57E-04	(0.39)	-1.37E-04	(0.39)

Figure 3: This figure plots marginal response probabilities for lagged values of BUSUVol and BISIVol, across the intervals used to classify the frequency distribution of D_k. The plots are referred to Bouygues. The central solid line represents the estimated marginal effect, while the two dashed lines define confidence intervals at 5%.

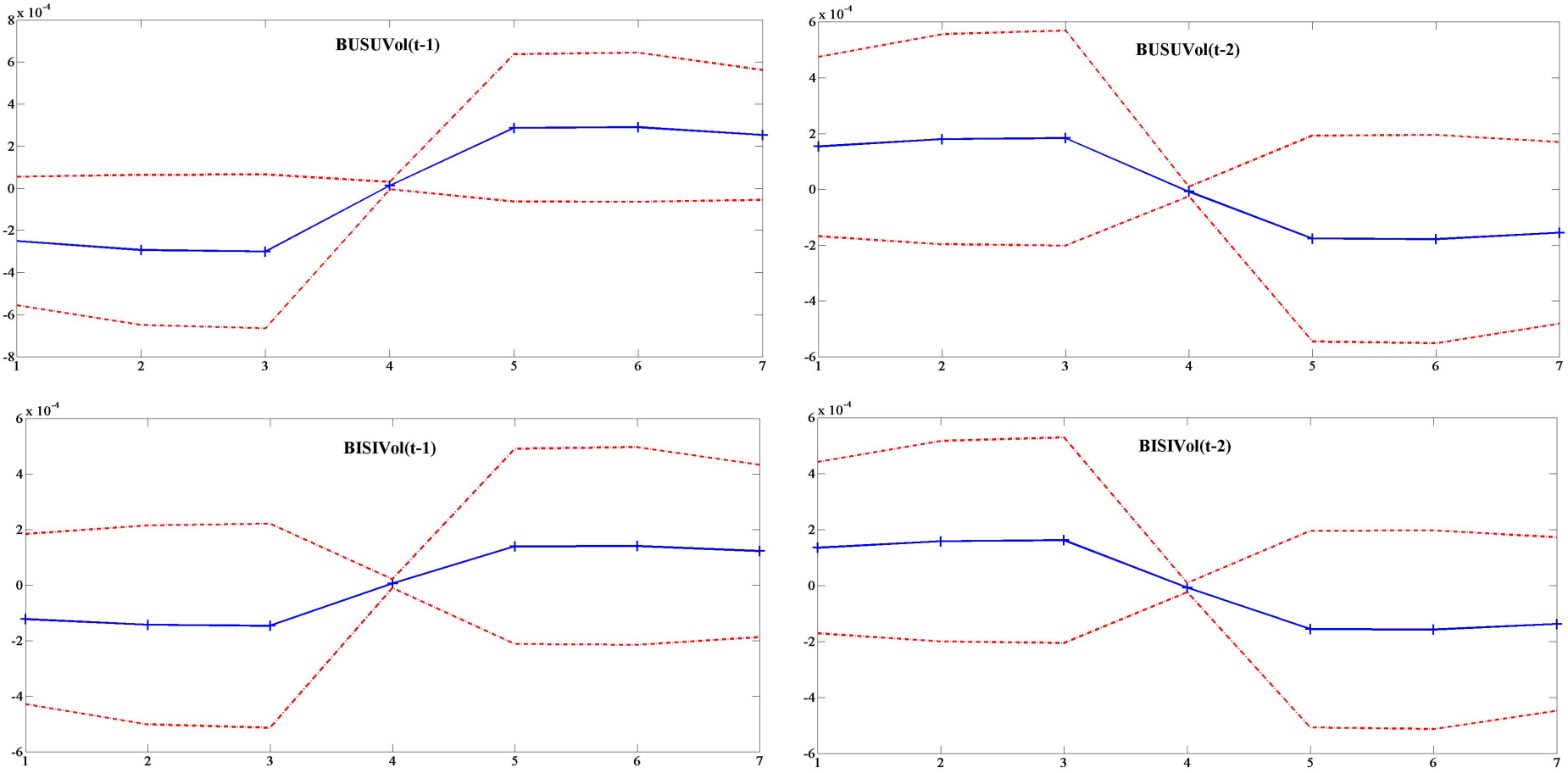

Figure 4: This figure plots marginal response probabilities for lagged values of BISUVol and BUSIVol across the intervals used to classify the frequency distribution of D_k. The plots are referred to Bouygues. The central solid line represents the estimated marginal effect, while the two dashed lines define confidence intervals at 5%.

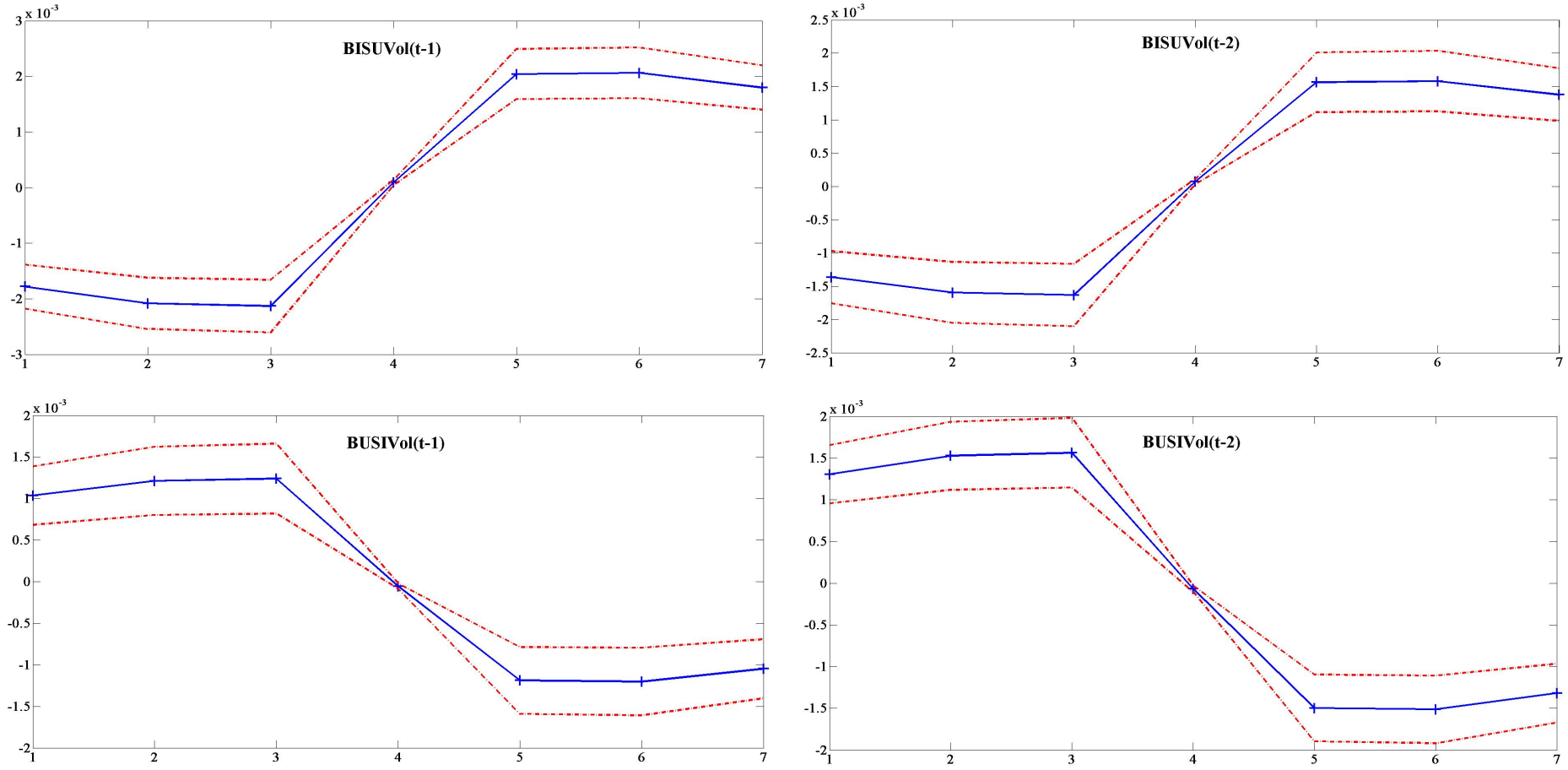

Table 9: This table provides the estimates of the bivariate probit model, for two stocks, Bouygues and Bnp. P-values are in parenthesis. The bottom lines displays goodness of fit statistics.

	Bouygues		Bnp	
	Buy	Sell	Buy	Sell
D_{open}	-9.81e-02	-1.46e-01	-6.40e-02	7.60e-02
	(0.00)	(0.00)	(0.00)	(0.00)
D_{lunch}	-4.00e-02	-4.24e-02	4.88e-02	2.03e-02
	(0.00)	(0.00)	(0.00)	(0.00)
D_{sp}	-1.36e-02	6.52e-02	-3.70e-02	-7.72e-04
	(0.21)	(0.00)	(0.00)	(0.91)
D_{clos}	-6.84e-02	1.05e-01	-3.10e-04	-7.47e-02
	(0.00)	(0.00)	(0.96)	(0.00)
δ_k	3.65e-06	3.32e-06	7.48e-06	1.08e-05
	(0.00)	(0.00)	(0.00)	(0.00)
Δt_k	-1.75e-03	-1.56e-03	-6.70e-03	-5.71e-03
	(0.00)	(0.00)	(0.00)	(0.00)
Volume	3.11e-06	2.20e-05	-1.71e-05	-3.62e-05
	(0.68)	(0.00)	(0.00)	(0.00)
D_{big}	1.13e-01	8.64e-02	7.61e-02	5.80e-02
	(0.00)	(0.00)	(0.00)	(0.00)
Init	1.79e-01	-1.15e-01	7.47e-02	-7.56e-02
	(0.00)	(0.00)	(0.00)	(0.00)
D_k	-3.21e-02	2.38e-02	-3.64e-02	3.13e-02
	(0.00)	(0.00)	(0.00)	(0.00)
SR	-7.90e+00	-1.50e-01	-2.05e+00	-1.73e+00
	(0.00)	(0.68)	(0.00)	(0.00)
Constant	9.61e-02	1.84e-01	-6.36e-03	6.49e-02
	(0.00)	(0.00)	(0.15)	(0.00)
QPS_{in}	0.48	0.47	0.49	0.48
QPS_{out}	0.49	0.51	0.49	0.49

Table 10: This table summarizes bivariate probit estimates over the whole sample. The first panel is referred to D_{b_k}, the second one is referred to D_{s_k}. In both cases, the first column exhibits the percentage of negative and significant estimates, the central one the quote of insignificant coefficients and the last one the fraction of positive and significant estimates.

	Negative Significant	Not significant	Positive Significant
D_{open}	66.67	28.21	5.13
D_{lunch}	28.21	28.21	43.59
D_{sp}	30.77	53.85	15.38
D_{clos}	38.46	28.21	33.33
δ_k	5.13	0.00	94.87
Δt_k	2.56	23.08	74.36
Volume	33.33	38.46	28.21
D_{big}	10.26	12.82	76.92
Init	0.00	0.00	100.00
D_k	0.00	0.00	100.00
SV	7.69	15.38	76.92

	Negative Significant	Not significant	Positive Significant
D_{open}	48.72	30.77	20.51
D_{lunch}	25.64	41.03	33.33
D_{sp}	15.38	51.28	33.33
D_{clos}	35.90	30.77	33.33
δ_k	2.56	5.13	92.31
Δt_k	0.00	15.38	84.62
Volume	38.46	33.33	28.21
D_{big}	7.69	7.69	84.62
Init	0.00	0.00	100.00
D_k	0.00	0.00	100.00
SV	7.69	17.95	74.36